基于多特征多分类器集成的专利自动分类研究*

贾杉杉[1] 刘 畅[2] 孙连英[3] 刘小安[1] 彭 涛[2]
[1](北京联合大学智慧城市学院 北京 100101)
[2](北京联合大学机器人学院 北京 100101)
[3](北京联合大学城市轨道交通与物流学院 北京 100101)

摘要:【目的】为了准确地给专利申请书分配IPC分类号, 本文提出一种基于多特征多分类器集成的专利自动分类方法。【方法】使用从专利申请书中提取的全词典TFIDF特征、信息增益词典TFIDF特征、段落向量特征、主题模型向量特征, 分别训练朴素贝叶斯、支持向量机、AdaBoost分类器, 以此构建特征–类别矩阵, 并结合F1权重矩阵集成, 获得最终IPC预测分类号。【结果】对2014年–2016年"发动机或泵"领域的10个小类进行分类, 使用Top Prediction、All Categories和Two Guesses三种评估方法得到准确率分别为: 78.9%、80.1%、91.2%。【局限】训练仅仅使用了2014年–2016年共三年的专利数据, 数据规模有限。【结论】在"发动机或泵"领域, 本文方法能够有效地提高专利文本分类的准确率。

关键词: 专利分类 段落向量 主题向量 分类器集成
分类号: G250

1 引 言

中国知识产权局研究发现[1], 知识资源和信息资源是最主要的智力资源, 尤其是专利信息这样基于创新、体现技术的资源。为了尽快找到和利用相关的专利信息, 需要对每一件专利按照其技术内容分配相应的专利分类号[2]。专利分类方法有很多, 其中使用最广泛的是国际专利分类(International Patent Classification, IPC)体系[3], 其几乎包括了与发明创造有关的全部知识领域, 中国、美国以及其他50多个国家和地区都在使用。

使用IPC分类体系进行专利分类难点众多, 主要有:

(1) 类别众多, 层次复杂, 最新的IPC分类体系有7万多个类别, 5个层级;

(2) 一件专利可被赋予不止一个分类号;

(3) 为了扩大专利受保护范围, 专利申请人对于专利申请的用词过于夸大;

(4) 类别之间相似度高, 对特征的表达能力要求高;

(5) 各个类别的专利数量严重不均衡, 给分类带来巨大压力。

目前, 专利审查员主要使用手工分类, 少量借助机器对专利进行分类。对于手工分类, 专利审查员需要逐篇阅读专利文献以确定分类号, 这样做效率低、费用高, 另外不同的人主观判断存在差别, 导致分类效果一致性较差[2]。近年来, 已有许多学者采用基于机器学习的方法对专利文本进行分类研究, 主要采用基于词的特征和单一分类器进行分类。然而这种方法并没有很好地解决专利文本分类这样复杂的文本分类任务。因此, 机器分类的准确率需要进一步提升, 以辅助

通讯作者: 彭涛, ORCID: 0000-0003-3533-9736, E-mail: pengtao@buu.edu.cn。
*本文系国家重点研发计划项目"公共安全风险防控与应急技术装备"(项目编号: 2016YFC0802107)和北京市教育委员会科技计划面上项目(项目编号: SQKM201411417013)的研究成果之一。

专利审查员的分类工作。

本文构建了 4 种特征：全词典 TFIDF 特征、信息增益词典 TFIDF 特征、段落向量特征、主题模型向量特征。使用朴素贝叶斯(或高斯–朴素贝叶斯)、支持向量机、AdaBoost 算法对 4 种特征分别训练得到 12 个分类器。从每一种特征对应的三个分类器中选取分类效果最好的分类器作为最优分类器。使用 4 个最优分类器的分类结果构建特征–类别矩阵，借助 F1 权重矩阵，对分类器进行集成，得到最终分类结果。

2　相关研究

国内外一直非常重视对专利文本的利用和研究。使用机器学习方法对专利文本按 IPC 分类号进行分类已经有近 20 年的历史。IPC 分类体系从上到下分为部、类、子类、组和子组 5 个级别。从 1971 年发布第一个版本，每 5 年更新一些子类以下的级别(组和子组)，最新 IPC 版本包含 7 万多个类别。

近年专利分类主要从三种角度展开研究工作：

(1) 将主题模型应用在特征中，使特征包含主题分布的信息。例如，Venugopalan 等[4]以自然语言处理为基础的分层技术，将太阳能光伏发电领域中 10 201 项专利的主题与现实世界中的类别/主题进行概率映射；廖列法等[5]提出使用 LDA 主题模型对专利进行分类，实验证明 LDA 主题模型比 KNN 方法的准确率高 10%以上。

(2) 神经网络和深度学习逐渐展露头角。马芳[2]使用标题和摘要，抽取 H 部 10 个相邻的小类 1 500 篇专利，利用径向基网络分类，准确率达到 72.2%。马双刚[6]选取计算机领域的发明专利，使用 SVM 对从自动编码器抽取的特征进行分类，准确率比传统的 SVM 提高 3.25%。

(3) 随着专利数据量逐年增加，并行化算法的研究得到重视。孔旗[7]提出 M3-SVM 算法，在大规模、不均衡数据集进行实验，精确率、召回率和 F1 测度三个指标都取得了比传统 SVM 更好的效果。

很多其他研究工作也共同丰富着专利分类方法。例如：刘桂锋等[3]提出基于概率超图的半监督的方法，在少量标记样本的情况下得到理想的分类效果。缪建明等[8]借助专利层次结构特点，仅使用摘要进行快速自动分类，大大提高了时效性。

在已有研究的基础上，本文构建 4 种特征：全部词的词典，以 TFIDF 为权重，构建代表全体词特征的 DIC_TFIDF 特征；通过信息增益算法构建信息增益词典，以 TFIDF 为权重，构建代表关键词信息的 IG_TFIDF 特征；训练段落向量，构建代表语义信息的 Document Vector 特征；训练 LDA，得到代表专利主题的 Topic Model Vector 特征。基于以上 4 组特征，分别训练 NB、SVM、AdaBoost 分类器。根据分类效果挑选最优分类器，构建特征–类别矩阵，并结合 F1 权重矩阵实现专利文本自动分类。

3　基于多特征分类器集成的专利文本分类

系统整体框架如图 1 所示，分为 4 个部分：预处理；构建 4 种分类器；选择最优分类器(框中 S、N、A 分别表示 SVM、NB、AdaBoost 分类器)；分类器集成。

3.1　预处理

从美国专利及商标局下载的数据是以周为单位的 XML 格式文件，解析成 TXT 格式文本，存入 MySQL 数据库。抽取发明名称、摘要、权利要求、详细说明、申请人、发明人等信息，对其进行分词、共指消解①、词干还原②等预处理。

3.2　4 组特征的构建

本文构建 4 组特征，分别如下：

(1) 针对预处理后的数据统计词典，每篇专利按照词典顺序构建词频矩阵。根据词频矩阵计算 TFIDF③值，得到代表全局信息的全词典的 TFIDF 权重的特征向量。

(2) 使用信息增益的方法，计算每个词对系统贡献值，从大到小排列。根据对比实验，选择前 4 351 个词。以此构建信息增益词典、词频矩阵，进而计算 TFIDF 值。根据以上信息得到代表关键词信息的信息增益词典的 TFIDF 权重特征向量。

①https: //stanfordnlp.github.io/CoreNLP/.
②http: //www.nltk.org/.
③http: //scikit-learn.org/stable/.

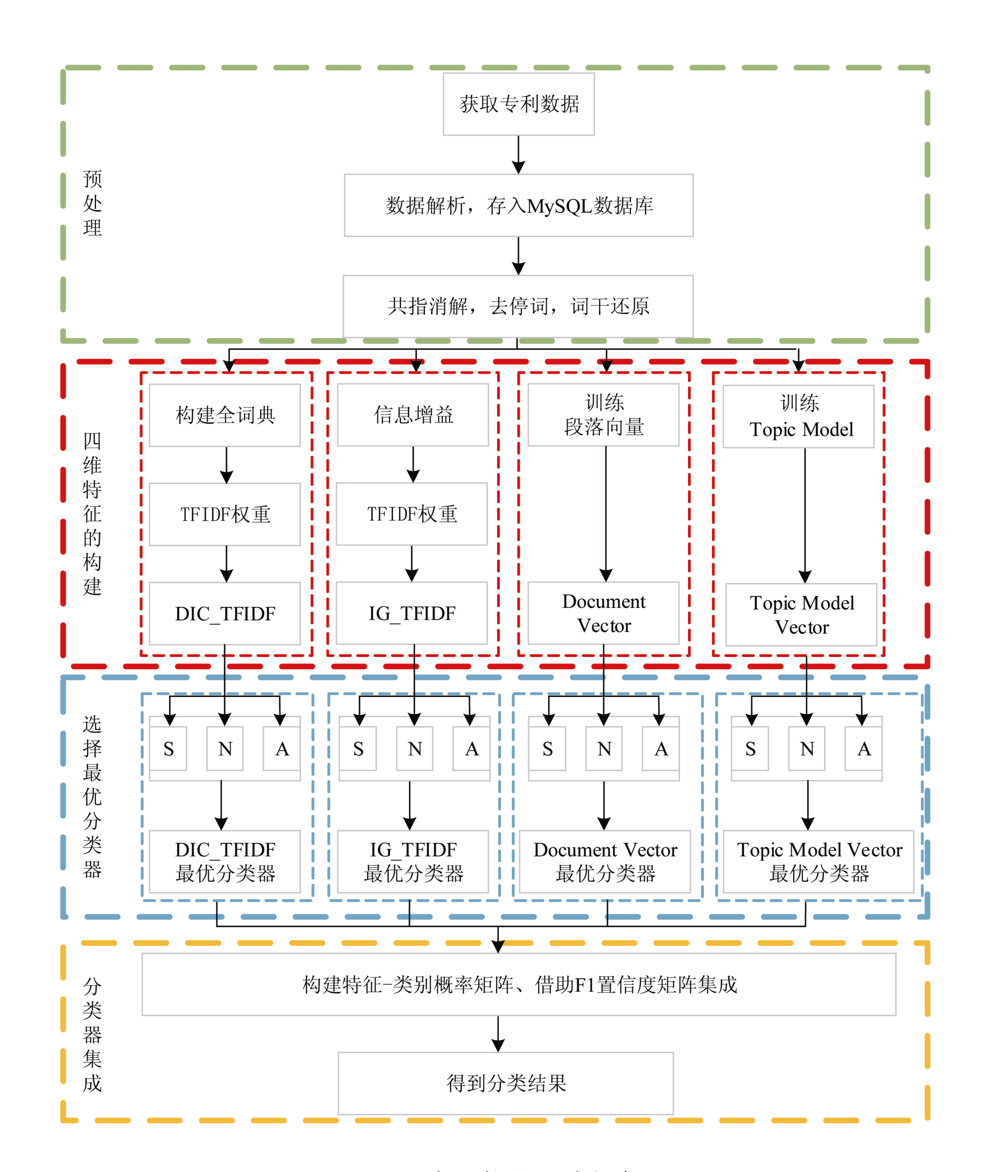


图 1　系统整体设计框架

(3) 由于词袋模型有两点主要的缺陷：丢失了词与词之间顺序的特征；忽略了词的语义信息。因此为每篇专利设计了包含语义的段落向量(Document Vector)。

Le 等[9]使用段落向量的分布式记忆模型(Distributed Memory Model of Paragraph Vector①)训练段落向量。段落向量模型在 Word2Vec 模型[10-11]基础上增加一个段落向量，同时将每个段落、每一个词映射为唯一的向量，如图 2 所示。

Classifier
Average/Concatenate
on
Paragraph Matrix
D
W
W
W
Paragraph ID
the
cat
sat

图 2　段落向量算法示意图[9]

①http: //radimrehurek.com/gensim/.

例如，训练两个段落“there are many animals in this room”和“the cat sat on table”，形成 50 维的段落向量和 50 维的词向量。训练阶段：首先初始化一个 2 行 50 列的 D 矩阵、12 行 50 列的 W 矩阵和 Softmax 的参数。当使用“the”、“cat”、“sat”预测下一个词时，从 D 中抽取出第二段对应的 1 个段落向量、W 中抽取出“the”、“cat”、“sat”对应的 3 个词向量，这 4 个向量求平均(或求和)，使用层次化的 Softmax 预测下一个词。整体使用随机梯度下降的方法训练，得到 D、W、Softmax 的参数。推理阶段：模型的 W 矩阵和 Softmax 的权重已固定，通过随机梯度下降获得新段落的段落向量。

(4) 由于每篇专利的主题不同，为专利文本设计了基于主题模型的主题向量。基于主题模型的 LDA 算法[12]，使用所有训集语料训练主题模型。由于 LDA 不仅限于“见过”的数据，可以通过训练好的主题模型得到一篇新文章的主题向量，并且相似主题的专利的主题向量也相似。

3.3 分类器集成

使用朴素贝叶斯、支持向量机和 AdaBoost 算法训练分类器，对 4 组特征分别训练。训练完成后，对测试集进行预测。在每一组特征所对应的三个分类器中，选择分类效果最好的作为该特征的最优分类器，共计 4 个。

设训练集有 N 个类别：$\{w_1,w_2,\cdots,w_N\}$，M 个最优分类器：$\{C_1,C_2,\cdots,C_M\}$。对于任意一个输入样本 x，令 $P_n^m(x)$ 表示使用最优分类器 C_m 计算该篇专利属于 w_n 的归一化后的值。全部的 $P_n^m(x)$ 组成 M 行 N 列的矩阵：特征–类别概率矩阵 $R(x)$ 如公式(1)所示，其中 $R(x)$ 的每列都对应一个类别，每行也对应一个特征的最优分类器：

$$R(x)=\begin{pmatrix} P_1^1(x) & P_2^1(x) & \cdots & P_N^1(x) \\ P_1^2(x) & P_2^2(x) & \cdots & P_N^2(x) \\ \cdots & \cdots & \cdots & \cdots \\ P_1^M(x) & P_2^M(x) & \cdots & P_N^M(x) \end{pmatrix} \quad (1)$$

首先，对于确定最优分类器对各类别预测结果的准确性这个问题，使用最优分类器对各类别进行分类的概率值归一化后结果 $P_n^m(x)$ 作为准确性的衡量标准。分类器的决策依据是：每个类别计算出相应的概率值越高，则属于该类别的可能性越高。

其次，对于多个分类器如何获得更好的集成效果的问题，不同最优分类器对不同类别的预测倾向不同，使用每个最优分类器对各类别的 F1 值作为细粒度权重。例如，最优分类器 1 认为样本 x 属于类别 1 的可能性为 $P_1^1(x)$，类推得到该样本属于每种类别的概率值。记概率最大值对应类别为预测值，统计测试集样本的预测值和真实值，得到每个分类器对每个类别的 F1 值。由于 F1 值同时兼顾召回率和精确率两个指标，常用来衡量分类器分类效果，因此使用 F1 值作为分类器集成的权重。

F1 权重矩阵如公式(2)所示。

$$F1=\begin{pmatrix} F_1^1 & F_1^2 & \cdots & F_1^M \\ F_2^1 & F_2^2 & \cdots & F_2^M \\ \cdots & \cdots & \cdots & \cdots \\ F_N^1 & F_N^2 & \cdots & F_N^M \end{pmatrix} \quad (2)$$

其中，F_N^M 表示最优分类器 M 将样本分到第 N 类别的 F1 值。

使用本文的多特征多分类器集成算法(Multi-Feature Multi-Classifier Integration, MFMCI)，累加每个分类器对其预测结果的概率值($P_n^m(x)$)与相应 F1 权重(F_n^m)的乘积。$F_n^m \times P_n^m(x)$ 表示有 F_n^m 的可能性，认为分类器 m 将样本 x 分到类别 n 是正确的，也是分类器 m 在多分类器集成中的贡献值。则 $S_n(x)$ 计算如公式(3)所示。

$$\begin{aligned} S_n(x) &= R(x)\cdot F1 \\ &= \sum_{m=1}^{M} F_n^m \times P_n^m(x), n=1,2,\cdots,N \end{aligned} \quad (3)$$

其中，M 为最优分类器的个数，N 为训练集中类别总数。

4 实验设计

4.1 实验语料和实验环境

专利数据下载自美国专利及商标局①(United States Patent and Trademark Office)，为 2014 年–2016

①United States Patent and Trademark Office: https://www.uspto.gov/.

年“发动机或泵”(Engine and Pump)领域的专利申请书。选择子类专利总数超过 800 篇的类别，从而得到 F01L、F01N、F02B、F02C、F02D、F02M、F03D、F04B、F04C、F04D 共 10 类，总计 8 000 篇专利数据，其中 5 500 篇作为训练集，2 500 篇作为测试集。

实验使用三台 CentOS7 64bit 操作系统、内存 16GB 的计算机。利用 Python 和 Java 语言，在 PyCharm 和 Eclipse 下编写程序并完成测试。整个实验阶段使用 sklearn、Stanford CoreNLP、gensim、NLTK 等依赖库。

4.2 评估方法和评价标准

以经典的准确率、召回率、F1 值和精确率作为评价标准。由于每篇专利的 IPC 分类号非唯一，所以使用三种不同的评估方法，如图 3 所示。Fall 等[13]也使用了这种评估方法。其中 Top Prediction 指分类器给出的第一预测值与第一真实值进行匹配; Two Guesses 是分类器前两个预测值与第一真实值进行匹配; All Categories 表示分类器给出第一预测值，与前三真实的分类号进行匹配。

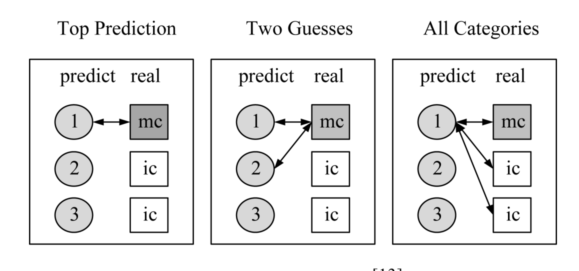


图 3 评估方法[13]

4.3 4 组特征的构建结果

预处理后所有训练语料的词作为词典(45 432 维)。构建每篇专利的基于全词典的词频向量矩阵，计算 TFIDF 权重，得到全词典 TFIDF 特征。

使用 Python 编写信息增益代码，计算每个词对整个系统的贡献值即信息增益值，得到 4 351 维，部分结果如表 1 所示。

表 1 部分词信息增益值

词(词干还原后)	信息增益值
Smoother	6.64337682087
vesda	6.64274815818
undamp	6.64274815818
engin	6.25488032208

训练段落向量时，分别以 50、100、150 和 200 维的段落向量进行实验，使用 SVM 对语料进行测试，最终选定 100 作为段落向量的维度。

使用 LDA 算法训练主题向量，需要根据数据特点确定主题的数量。由于样本共有 10 个类别，因此分别设计 10、12、15、18 和 20 个主题的对比实验，根据实验结果最终选择 15 作为主题向量的主题个数。

4.4 选择最优分类器

每个特征训练三个分类器，分别是：朴素贝叶斯(NB)或高斯–朴素贝叶斯(Gaussian-NB)、支持向量机(SVM)和 AdaBoost。由于全字典特征维数太高，只得到基于朴素贝叶斯分类器的分类效果。所有分类器表现效果如表 2 所示。

表 2 各分类器不同特征下表现效果

分类算法	特征	评估方法	准确率	F1 值	召回率	精确率
NB	**全字典 TFIDF**	**Top Prediction**	**71.4%**	**71.1%**	**71.4%**	**72.3%**
NB	信息增益 TFIDF	Top Prediction	43.9%	44.7%	43.9%	46.1%
SVM	信息增益 TFIDF	Top Prediction	64.6%	64.4%	64.6%	68.0%
AdaBoost	**信息增益 TFIDF**	**Top Prediction**	**71.7%**	**71.9%**	**71.7%**	**72.9%**
Gaussian-NB	段落向量	Top Prediction	23.3%	21.4%	23.3%	24.3%
SVM	**段落向量**	**Top Prediction**	**48.4%**	**48.2%**	**48.4%**	**48.7%**
AdabBoost	段落向量	Top Prediction	23.6%	23.6%	23.6%	24.1%
Gaussian-NB	主题向量	Top Prediction	39.7%	38.3%	39.7%	39.6%
SVM	**主题向量**	**Top Prediction**	**41.7%**	**40.4%**	**41.7%**	**42.2%**
AdaBoost	主题向量	Top Prediction	41.6%	40.8%	41.6%	40.7%

根据实验结果比对，每组选择分类效果最好的分类器作为该特征的最优分类器。全字典 TFIDF 选择朴素贝叶斯分类器，信息增益 TFIDF 特征选择 AdaBoost 分类器，段落向量特征选择 SVM 分类器，主题向量特征选择 SVM 分类器。

5 实验结果与分析

5.1 实验结果

不同特征组合与分类器集成的实验结果，如表 3 所示。

表 3 不同特征组合与分类器集成的实验结果

实验	评估方法	特征	算法	准确率	F1 值	召回率	精确率
1	All Categories	Ⅰ全字典 TFIDF	NB	73.6%	73.5%	73.6%	74.6%
2	All Categories	Ⅱ信息增益 TFIDF	AdaBoost	74.0%	73.0%	74.0%	76.7%
3	All Categories	Ⅲ段落向量	SVM	49.4%	49.1%	49.4%	49.6%
4	All Categories	Ⅳ主题向量	SVM	42.0%	41.3%	42.0%	41.6%
5	All Categories	Ⅱ、Ⅲ、Ⅳ直接拼接	Gaussian-NB	31.2%	30.8%	31.2%	31.6%
6	All Categories	Ⅲ、Ⅳ特征直接拼接	SVM	34.4%	33.2%	34.4%	34.1%
7	All Categories	Ⅰ、Ⅱ、Ⅲ、Ⅳ	投票	72.2%	73.5%	72.2%	74.0%
8	All Categories	Ⅱ、Ⅲ、Ⅳ	MFMCI	54.1%	52.0%	54.1%	56.6%
9	All Categories	Ⅰ、Ⅲ、Ⅳ	MFMCI	79.4%	78.8%	79.4%	81.7%
10	**All Categories**	Ⅰ、Ⅱ、Ⅲ、Ⅳ	**MFMCI**	**80.1%**	**79.5%**	**80.1%**	**82.4%**
11	Top Prediction	Ⅰ全字典 TFIDF	NB	71.4%	71.1%	71.4%	72.3%
12	Top Prediction	Ⅱ信息增益 TFIDF	AdaBoost	71.7%	71.9%	71.7%	72.9%
13	Top Prediction	Ⅲ段落向量	SVM	48.4%	48.2%	48.4%	48.7%
14	Top Prediction	Ⅳ主题向量	SVM	41.7%	40.4%	41.7%	42.2%
15	Top Prediction	Ⅱ、Ⅲ、Ⅳ直接拼接	Gaussian-NB	31.2%	30.8%	31.2%	31.6%
16	**Top Prediction**	Ⅰ、Ⅱ、Ⅲ、Ⅳ	**MFMCI**	**78.9%**	**78.2%**	**78.9%**	**81.2%**
17	Two Guesses	Ⅰ全字典 TFIDF	NB	88.1%	88.1%	88.1%	88.4%
18	Two Guesses	Ⅱ信息增益 TFIDF	AdaBoost	89.4%	89.2%	89.4%	89.8%
19	Two Guesses	Ⅲ段落向量	SVM	68.6%	68.5%	68.6%	68.7%
20	Two Guesses	Ⅳ主题向量	SVM	61.8%	61.4%	61.8%	61.9%
21	**Two Guesses**	Ⅰ、Ⅱ、Ⅲ、Ⅳ	**MFMCI**	**91.2%**	**91.0%**	**91.2%**	**91.7%**

5.2 实验结果分析

从单特征单分类器的实验结果(1-4、11-14、17-20)得到综合实验效果是：信息增益 TFIDF 特征最优分类器>全字典 TFIDF 特征最优分类器>段落向量特征最优分类器>主题向量特征最优分类器。由此可见，对于单个特征而言，信息增益方法选择的特征包含信息量最大，主题向量特征包含的信息量最少。全字典特征总共有 45 432 维，信息增益 4 351 维，段落向量 100 维，主题向量仅仅 15 维，特征维数的差距巨大，是造成信息包含量不同的原因之一。然而，并不是特征维数越大，分类效果一定更好，段落向量特征选择 100 维时的分类效果比 200 维效果好。另外，由于特征维度的不同，不同算法的分类效果也不同。例如，全字典 45 432 维特征，使用贝叶斯分类效果最好。维度过高对于计算复杂度很高的 SVM 来说很难达到好的效果。因此设计每种特征分别训练三个不同算法的分类器，从中选择分类效果好的作为最优分类器。

由表 3 可以看出，采用本文算法在三种评估方法下都取得最好的准确率，All Categories 达到 80.1%，Top Prediction 达到 78.9%，Two Guesses 达到 91.2%。由此可以得到 4 点结论：

(1) 4 种特征包含的信息量远大于单个特征的信息。4 种特征包含全局单词的信息(全词典)、关键词信息(信息增益)、关于语义的信息(段落向量)、篇章主题的信息(主题向量)。多特征使得对专利信息的描述更具全局化和立体化，效果比片面的局部的特征分类效

果好。

(2) 信息增益 TFIDF 特征是从全字典中提取出的关键词特征。从实验 1-2、11-12、17-18 看到，信息增益 TFIDF 最优分类器的分类效果比全词典好。但从实验 8-10 结果分析，全字典特征更能代表全局信息，而信息增益则相对局部。此外，它们对一篇专利的预测概率以及集成时的 F1 值完全不同，如表 4 所示。因此在集成过程中，全词典的全局信息是对信息增益的局部信息的补充。所以，两者与其他两种特征的最优分类器的集成效果会远远好于其中一者与其他两种特征最优分类器的集成效果。

表 4　全词典 TFIDF 最优分类器和信息增益 TFIDF 最优分类器的区别

IPC 分类号	F1 值		对某篇专利预测概率值	
	全词典 TFIDF 最优分类器	信息增益 TFIDF 最优分类器	全词典 TFIDF 最优分类器	信息增益 TFIDF 最优分类器
F01L	86.1%	83.4%	66.5%	11.342%
F01N	78.1%	74.2%	0.6%	10.001%
F02B	59.8%	53.8%	10.9%	10.019%
F02C	76.0%	87.2%	0.6%	9.588%
F02D	67.1%	58.3%	9.6%	10.022%
F02M	57.7%	50.6%	3.2%	10.006%
F03D	94.1%	96.4%	0.3%	9.035%
F04B	72.6%	75.6%	7.1%	10.004%
F04C	74.7%	77.2%	1.0%	9.992%
F04D	69.0%	62.7%	0.3%	9.989%

(3) 并不是 4 种特征随便结合到一起就可以提高分类效果，从表 3 中实验 5 和实验 6 得到，将特征直接拼接，分类效果急剧下降。这种做法是无效的，只有将特征有机结合才能更好地发挥各自的优势。

(4) 本文算法不是对多个特征分类器的简单投票，对比实验 7 和实验 10 发现，想要有机结合多个特征分类器，需要抓住各自特征分类的优势，在有优势的地方加大权重，在劣势的地方给予低的权重。结合 F1 权重矩阵与特征–类别概率矩阵，最终得到更好的分类效果。

马芳[2]使用径向基神经网络对专利自动分类(记为 RBFNN)，本文工作与其都是将专利分到小类级别的 10 个类，相对具有可比性。本文比马芳的使用径向基神经网络的分类效果好，结果如表 5 所示。

表 5　本文与其他工作效果对比

算法	标准	准确率	F1 值	召回率
RBFNN(径向基网络)	Top Prediction	72.2%	70.7%	71.0%
MFMCI(本文算法)	Top Prediction	78.9%	78.2%	78.9%
MFMCI(本文算法)	All Categories	80.1%	79.5%	80.1%
MFMCI(本文算法)	Two Guesses	91.2%	91.0%	91.2%

MFMCI 对各个类别的预测效果如表 6 所示。

表 6　MFMCI 对各个类别的预测效果

IPC 分类号	F1 值	召回率	精确率
F01L	**86.21%**	98.8%	76.5%
F01N	**85.60%**	86.8%	84.4%
F02B	66.67%	53.2%	89.3%
F02C	**82.93%**	81.6%	84.3%
F02D	73.31%	95.6%	59.5%
F02M	64.99%	51.6%	87.8%
F03D	**91.01%**	97.2%	85.6%
F04B	79.29%	71.2%	89.5%
F04C	**84.54%**	90.8%	79.1%
F04D	**80.87%**	74.4%	88.6%

从表 6 中可以看出，MFMCI 对 F01L、F01N、F02C、F03D、F04C、F04D 等类的分类效果比较好，F1 值都超过了 80%。而对 F02B、F02D、F02M、F04B 等类的分类效果不好，究其原因有以下两点：

(1) 这 10 个类别有非常多的交叉和相似之处，例如最新的 IPC 国际专利分类标准记载：F02B：活塞式内燃机；一般燃烧发动机(其循环操作阀入 F01L；内燃机润滑入 F01M；其气流消音器或排气装置入 F01N；内燃机的冷却入 F01P；燃气轮机入 F02C；利用燃烧生成物的发动机装置入 F02C, F02G)。F02B 代表“活塞式内燃机；一般燃烧发动机”。但是如果是燃气轮机就要转入 F02C，若专利是利用燃烧生成物的发动机装置也被转入 F02C。

(2) 部分专利申请人为了扩大自己专利的权利范围，故意在申请书中扩大用词。针对这种情况，机器很难仅仅借助文本对专利进行有效分类，在以后的工作中，可以考虑引入专利申请中的图像的特征来提高效果。

6　结　语

本文提出一种多特征多分类器集成的专利自动分类算法。该方法以全局词特征(全词典 TFIDF 特征)、

关键词特征(信息增益 TFIDF 特征)、语义特征(篇章向量特征)、主题特征(主题向量特征), 分别训练属于每个特征最好的分类器作为最优分类器, 构建特征–类别概率矩阵, 结合 F1 权重矩阵, 对发动机或泵领域的 10 个子类进行分类。与单特征、直接串联特征、多特征分类器直接投票、以及马芳[2]的径向基神经网络方法的分类效果进行对比, 能够取得较好结果。

本文的不足以及未来工作主要有: 对于专利申请人故意扩大用词的情况可以借助专利申请中的图像辅助分类; 本文使用了近三年的专利, 数量不足以支撑完成组和子组级别的分类。未来研究可以使用 2001 年至今近 17 年的专利数据, 结合分布式分类算法, 对专利进行更深层次的分类, 以进一步提高分类准确率。

参考文献:

[1] 蔡虹, 蒋仁爱, 吴凯. 知识产权保护对中国技术进步的贡献研究[J]. 系统管理学报, 2015, 24(3): 314-320. (Cai Hong, Jiang Renai, Wu Kai. Contribution of Intellectual Property Protection to the Technological Progresses in China [J]. Journal of Systems & Management, 2015, 24(3): 314-320.)

[2] 马芳. 基于 RBFNN 的专利自动分类研究[J]. 现代图书情报技术, 2011(12): 58-63. (Ma Fang. Research of Patent Automatic Classification Based on RBFNN [J]. New Technology of Library and Information Service, 2011(12): 58-63.)

[3] 刘桂锋, 汪满容, 刘海军. 基于概率超图半监督学习的专利文本分类方法研究[J]. 情报杂志, 2016 , 35(9) : 187-191, 173. (Liu Guifeng, Wang Manrong, Liu Haijun. Probabilistic Hypergraph Based Semi-supervised Learning Method for Patent Document Categorization[J]. Journal of Intelligence, 2016, 35(9): 187-191, 173.)

[4] Venugopalan S, Rai V. Topic Based Classification and Pattern Identification in Patents[J]. Technological Forecasting and Social Change, 2015, 94: 236-250.

[5] 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报, 2017, 37(3): 35-39. (Liao Liefa, Le Fugang, Zhu Yalan. The Application of LDA Model in Patent Text Classification[J]. Journal of Modern Information, 2017, 37(3): 35-39.)

[6] 马双刚. 基于深度学习理论与方法的中文专利自动分类研究[D]. 镇江: 江苏大学, 2016. (Ma Shuanggang. The Study of Automatic Chinese Patent Classification Based on Deep Learning Theory and Method [D]. Zhenjiang: Jiangsu University, 2016.)

[7] 孔旗. 基于并行机器学习的大规模专利分类[D]. 上海: 上海交通大学, 2011. (Kong Qi. Large-scale Patent Classification Based on Parallel Machine Learning [D]. Shanghai: Shanghai Jiaotong University, 2011.)

[8] 缪建明, 贾广威, 张运良. 基于摘要文本的专利快速自动分类方法[J]. 情报理论与实践, 2016, 39(8): 103-105, 91. (Miu Jianming, Jia Guangwei, Zhang Yunliang. The Rapid Automatic Categorization of Patent Based on Abstract Text [J]. Information Studies: Theory & Application, 2016, 39(8): 103-105, 91.)

[9] Le Q V, Mikolov T. Distributed Representations of Sentences and Document[OL]. arXiv Preprint, arXiv: 1405.4053.

[10] Mikolov T. Statistical Language Models Based on Neural Networks[D]. Brno University of Technology, 2012.

[11] Turian J, Ratinov L, Bengio Y. Word Representations: A Simple and General Method for Semi-supervised Learning [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 384-394.

[12] Rosen-Zvi M, Griffiths M, Steyvers M, et al. The Author-topic Model for Authors and Documents[C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. 2012: 487-494.

[13] Fall C J, Törcsvári A, Benzineb K, et al. Automated Categorization in the International Patent Classification[J] . ACM SIGIR Forum, 2003, 37 (1): 10-25.

作者贡献声明:

贾杉杉: 提出研究思路, 设计研究方案, 进行实验, 起草论文;
彭涛: 采集、清洗和分析数据;
刘畅, 孙连英, 刘小安, 彭涛: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: pengtao@buu.edu.cn。

[1] 贾杉杉, 刘畅, 孙连英, 刘小安, 彭涛. 内在数据详细信息. doc. 研究方法和工具直接描述、样本数据下载地址.

[2] 贾杉杉, 刘畅, 孙连英, 刘小安, 彭涛. 附件–表 2 和表 3 原始实验结果.txt. 直接研究结果数据.

收稿日期: 2017-05-31
收修改稿日期: 2017-08-15

Patent Classification Based on Multi-feature and Multi-classifier Integration

Jia Shanshan[1] Liu Chang[2] Sun Lianying[3] Liu Xiaoan[1] Peng Tao[2]
[1](College of Intellectualized City, Beijing Union University, Beijing 100101, China)
[2] (College of Robotics, Beijing Union University, Beijing 100101, China)
[3] (College of Urban Rail Transit and Logistics, Beijing Union University, Beijing 100101, China)

Abstract: [**Objective**] This paper aims to automatically allocate correct IPC to patent applications with the help of multi-feature and multi-classifier integration method. [**Methods**] First, we extracted the TFIDF features of all dictionaries and information gains, as well as the vector features of document and topic models from patent applications. Then, we used the collected data to train the NB, SVM, and AdaBoost classifiers. Finally, we established the feature-class matrix and predicted the final IPC with the F1 weight matrix. [**Results**] We examined our new method with 10 patent classes from 2014 to 2016 in the field of engine and pump. The accuracy of top prediction, all categories, and two guesses were 78.9%, 80.1% and 91.2% respectively. [**Limitations**] The size of training corpus is limited, which only includes 3 years patent data. [**Conclusions**] The proposed method could effectively improve the accuracy of patent classification in the field of engine and pump.

Keywords: Patent Classification Document Vector Topic Model Vector Classifier Integration

---

国家纳米科学中心、中国科学院文献情报中心、施普林格·自然联合推出 中国纳米科学与技术发展白皮书

北京国际会议中心举办的 2017 年中国国际纳米科学技术会议上，国家纳米科学中心、中国科学院文献情报中心和施普林格·自然集团(Springer Nature)联合发布了《国之大器始于毫末——中国纳米科学与技术发展状况概览》中英文白皮书。

中国投入进行纳米科研已有数十年时间，已经成为当今世界纳米科学与技术进步重要的贡献者，部分基础研究居国际领先水平，中国纳米科技应用研究与成果转化的成效也已初具规模。这些都与中国在纳米科技领域的持续投入密切相关。中国纳米科技研究正在向原创性突破转变，并更加关注纳米科技的产业化应用。

白皮书分别从原创论文数量、Nano 数据库和专利产出这三个方面，将中国与世界其他主要纳米科研强国进行了对比，揭示了中国纳米科研的优势与发展特点。白皮书还通过业内专家访谈，探讨了中国纳米科学的发展前景和未来面临的挑战。

中国科学院院长、党组书记白春礼指出，从计量学角度在对纳米科技成果分析的基础上，进一步关注纳米专利技术的应用情况，关注纳米研发的投入成效，更深入地揭示和把握纳米科技的发展态势。

(1) 中国纳米科技论文：产出数量和质量均有大幅提升；

(2) Nano 数据库彰显中国纳米研究的优势与侧重；

(3) 中国纳米专利：数量全球第一，但多为本国专利。

白皮书指出，科研产出和专利申请数量上的迅速增长，都描绘出中国纳米科学美好的发展前景。不论是传统的强项学科，还是新兴领域，中国纳米科学都表现出巨大的潜力。

(本刊讯)